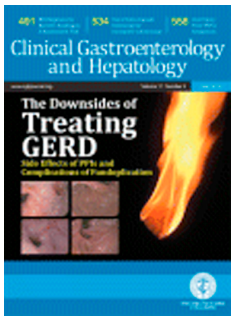
DOI:
10.1016/j.cgh.2019.05.061

Document Version
Peer reviewed version

Accepted Manuscript



Genetic and Inflammatory Biomarkers Classify Small Intestine Inflammation in Asymptomatic First-degree Relatives of Patients With Crohn's Disease

Kirstin M. Taylor, Ken B. Hanscombe, Natalie J. Prescott, Raquel Iniesta, Matthew Traylor, Nicola S. Taylor, Steven Fong, Nicholas Powell, Peter M. Irving, Simon H. Anderson, Christopher G. Mathew, Cathryn M. Lewis, Jeremy D. Sanderson

PII: S1542-3565(19)30643-3
DOI: https://doi.org/10.1016/j.cgh.2019.05.061
Reference: YJCGH 56569

To appear in: *Clinical Gastroenterology and Hepatology*
Accepted Date: 29 May 2019

Please cite this article as: Taylor KM, Hanscombe KB, Prescott NJ, Iniesta R, Traylor M, Taylor NS, Fong S, Powell N, Irving PM, Anderson SH, Mathew CG, Lewis CM, Sanderson JD, Genetic and Inflammatory Biomarkers Classify Small Intestine Inflammation in Asymptomatic First-degree Relatives of Patients With Crohn's Disease, *Clinical Gastroenterology and Hepatology* (2019), doi: https://doi.org/10.1016/j.cgh.2019.05.061.

Manuscript Number: CGH 18-01753

Title: Genetic and Inflammatory Biomarkers Classify Small Intestine Inflammation in Asymptomatic First-degree Relatives of Patients With Crohn's Disease

Short title: Pre-symptomatic detection of Crohn's disease

*Kirstin M. Taylor[1,2], *Ken B. Hanscombe[1], *Natalie J. Prescott[1], Raquel Iniesta[1,3], Matthew Traylor[4], Nicola S. Taylor[5], Steven Fong[2], Nicholas Powell[2], Peter M. Irving[2], Simon H. Anderson[2], Christopher G. Mathew[1,6], Cathryn M. Lewis[1], Jeremy D. Sanderson[2]

1. Department of Medical and Molecular Genetics, King's College London, London, United Kingdom
2. Department of Gastroenterology, Guy's & St Thomas' NHS Foundation Trust, St Thomas' Hospital, London, United Kingdom
3. Department of Biostatistics & Health Informatics, King's College London,
4. Department of Clinical Neurosciences, University of Cambridge, Cambridge, United Kingdom
5. Department of Gastroenterology, Southampton General Hospital, Southampton, United Kingdom
6. Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

*authors contributed equally to this work

Grant support: This study was undertaken as part of a Guy's & St Thomas' Charity funded clinical research fellowship (R080522). We also acknowledge support from the National Institutes of Health Research Biomedical Research Centres at Guy's & St Thomas' NHS Foundation Trust, and at South London and Maudsley NHS Foundation Trust, in partnership with King's College London. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abbreviations: AIC – Akaike's information criterion; ASCA – anti-saccharomyces cerevisiae antibodies; AUC – area under the curve; BIC – Bayesian information criterion; CD – Crohn's disease; CECDAI – Capsule Endoscopy Crohn's Disease Activity Index ; CEU – Utah residents with Northern and Western European ancestry from the CEPH (Centre d'Etude du Polymophisme Humain); CI – confidence interval; DNA – deoxyribonucleic acid; ELISA – enzyme-linked immunosorbent assay; FC – fecal calprotectin; FDR – first degree relative; GWAS – genome wide association study; Hs-CRP – high sensitivity C reactive protein; IBD – inflammatory bowel disease; NOD2 – nucleotide oligomerisation domain 2 gene; OR – odds ratio; QC – quality control; ROC – receiver-operator characteristic; RR – relative risk; SI – small intestine; SNP – single nucleotide polymorphism; VCE – video capsule endoscopy;

Correspondence: Natalie J Prescott, Department of Medical and Molecular Genetics, Kin's College London, 7th Floor Tower Wing, Guy's Hospital, London, SE1 9RT. email: Natalie.prescott@kcl.ac.uk

Disclosures: all authors declare that they are no conflict of interests to disclose.

Authors' contributions: KMT: study design; participant recruitment and follow-up; acquisition of samples; DNA extraction and QC; performance, reading and scoring of capsule endoscopies; data collection, analysis and interpretation; drafting of the manuscript; critical revision of the manuscript for important intellectual content
KBH: data analysis; analysis and interpretation of genetics data; advanced statistical analysis including explanatory and predictive modelling; drafting of the manuscript; critical revision of the manuscript for important intellectual content
RA: interpretation of data; advanced statistical analysis; critical revision of the manuscript for important intellectual content
MT: interpretation of data; advanced statistical analysis; critical revision of the manuscript for important intellectual content
NST: reading and scoring of capsule endoscopies; interpretation of data; critical revision of the manuscript for important intellectual content
NP: analysis and interpretation of biomarker data; critical revision of the manuscript for important intellectual content
PMI: participant recruitment and follow-up; acquisition of data; critical revision of the manuscript for important intellectual content
SHA: study concept and design; reading and scoring of capsule endoscopies; critical revision of the manuscript for important intellectual content; study supervision
NJP: study concept and design; assistance with DNA extraction and QC; assistance with analysis and interpretation of genetics data; critical revision of the manuscript for important intellectual content; study supervision
CGM: study concept and design; assistance with analysis and interpretation of genetics data; critical revision of the manuscript for important intellectual content; study supervision
CML: study concept and design; assistance with analysis and interpretation of genetics data; assistance with advanced statistical analysis; critical revision of the manuscript for important intellectual content; study supervision
JDS: study concept and design; analysis and interpretation of data; critical revision of the manuscript for important intellectual content; obtained funding; study supervision

Abstract:

Background & Aims: Relatives of individuals with Crohn's disease (CD) carry CD-associated genetic variants and are often exposed to environmental factors that increase their risk for this disease. We aimed to estimate the utility of genotype, smoking status, family history, and other biomarkers can be used to calculate risk in asymptomatic first-degree relatives of patients with CD.

Methods: We recruited 480 healthy first-degree relatives (full siblings, offspring or parents) of patients with CD through the Guy's and St Thomas' NHS Foundation Trust and from members of Crohn's and Colitis, United Kingdom. DNA samples were genotyped using the Immunochip. We calculated a risk score for 454 participants, based on 72 genetic variants associated with CD, family history, and smoking history. Participants were assigned to highest and lowest risk score quartiles. We assessed pre-symptomatic inflammation by capsule endoscopy and measured 22 markers of inflammation in stool and serum samples (reference standard). Two machine-learning classifiers (elastic net and random forest) were used to assess the ability of the risk factors and biomarkers to identify participants with small intestinal inflammation in the same dataset.

Results: The machine-learning classifiers identified participants with pre-symptomatic intestinal inflammation: elastic net (area under the curve, 0.80; 95% CI, 0.62–0.98) and random forest (area under the curve, 0.87; 95% CI, 0.75–1.00). The elastic net method identified 3 variables that can be used to calculate odds for intestinal inflammation: combined family history of CD (odds ratio, 1.31), genetic risk score (odds ratio, 1.14), and fecal level of calprotectin (odds ratio, 1.04). These same 3 variables were among the 5 factors associated with intestinal inflammation in the random forest model.

Conclusion: Using machine learning classifiers, we found that genetic variants associated with CD, family history, and fecal level of calprotectin together identify individuals with pre-symptomatic intestinal inflammation who are therefore at risk for CD. A tool for detecting people at risk for CD before they develop symptoms would help identify the individuals most likely to benefit from early intervention.

KEY WORDS:

What You Need to Know

Background: Relatives of individuals with Crohn's disease (CD) carry CD-associated genetic variants and are often exposed to environmental factors that increase their risk for this disease. We investigated whether genotype, smoking status, family history, and other biomarkers can be used to calculate risk in asymptomatic first-degree relatives of patients with CD.

Findings: Using machine learning classifiers, we found that genetic variants associated with CD, family history, and fecal level of calprotectin together identify individuals with pre-symptomatic intestinal inflammation who are therefore at risk for CD.

Implications for patient care: A tool for detecting people at risk for CD before they develop symptoms would help identify the individuals most likely to benefit from early intervention.

Introduction

Inflammatory bowel disease (IBD), comprising Crohn's disease (CD) and ulcerative colitis (UC), is a chronic inflammatory condition of the gastrointestinal tract associated with significant morbidity. Family members of CD patients are at increased risk for developing the disease with estimates of sibling risk ranging from 15-42 times greater than the general population[1].

Although CD pathogenesis remains incompletely understood, the role of genetic risk factors is well established. Twin studies estimate that the heritability (proportion of trait variance explained by genetic factors), is 0.756. To date, there are over 240 genetic variants or single nucleotide polymorphisms (SNPs) robustly associated with IBD through genome-wide association studies (GWAS)[2] which account for approximately half of the heritability estimate for CD[3]. A recent study of first-degree relatives (FDRs) of patients with IBD showed that they are enriched for IBD-associated risk loci[4].

Familial clustering of disease is likely the result of shared environmental factors, as well as genetic risk factors. The increasing incidence of the disease among populations with historically lower rates, and those migrating from regions with low incidence rates to regions with higher rates, is consistent with a substantial environmental risk component to CD[5]. Many lifestyle-related factors have been implicated, including stress, sedentary lifestyle, western diet, poor sleep, and tobacco use[6,7]. Smoking is the best-studied environmental risk factor with recent estimates suggesting a nearly two-fold increase risk for CD[8].

Asymptomatic FDRs may display phenotypic features in common with CD patients including altered intestinal permeability, positive serological antimicrobial markers, disordered innate and acquired

immunity, faecal dysbiosis, and elevated faecal calprotectin (FC)[9]. Overt small intestinal (SI) inflammation has been described in FDRs who have undergone ileocolonoscopy[10], intestinal ultrasound[11]and video capsule endoscopy (VCE)[12].

Evidence of increased risk factors for CD in FDRs raises the possibility of predicting those at risk of developing the disease[9]. We hypothesized that higher genetic risk in FDRs compared to healthy controls[4], elevated FC levels[9], and smoking, taken together could provide sufficient information to detect at-risk individuals before the development of overt symptoms. This has the potential to enable early interventions to alter or halt aberrant immune and inflammatory responses, and perhaps, ultimately, prevent disease. In this study we aimed to assess the clinical utility of a disease risk model to predict the presence of SI inflammation as detected by VCE. We applied two machine-learning methods to risk factors including genetic variants, smoking status, family history of disease and 22 inflammatory biomarkers, and assessed the predictive ability of the derived models.

Materials and Methods

Participants

We recruited 480 healthy FDRs (full siblings, offspring or parents) through CD patients attending the IBD service at Guy's & St Thomas' NHS Foundation Trust (GSTT) and from members of Crohn's & Colitis UK (a charity supporting patients with IBD). CD diagnosis in the probands was confirmed by their gastroenterologist or general practitioner. FDRs provided information regarding family history of IBD, medical history (including gastrointestinal symptoms), smoking status, medications, allergies, primary care provider and ethnicity. In keeping with the population of the CD GWAS meta-analysis[13], only FDRs of European ancestry were included. Further inclusion criteria were: ability to give written

informed consent, age 18-55 years, and absence of gastrointestinal symptoms. The exclusion criteria were: a previous diagnosis of IBD, irritable bowel syndrome, or other major co-morbidity, pregnancy, major bowel surgery, or the use of non-steroidal anti-inflammatory drugs in the 4 weeks prior to capsule endoscopy (low-dose aspirin excluded).

DNA collection and genotyping

Saliva sampling kits (Oragene™ DNA OG-500) were sent to 480 FDRs that met the inclusion criteria. DNA samples were genotyped on the Immunochip[14] which included 72 known significant CD SNPs at the time of study design (Supplementary methods). Genotypes for these 72 variants were extracted for genetic risk profiling and underwent standard quality control (QC) in PLINK[13,15] (Supplementary methods).

Calculation of high and low risk groups

We generated a combined CD risk score in 454 FDRs based on genotype (**Table S1**)[13] and smoking status (current, ex, never-smoker)[16] using the R package REGENT[17]. Summary results were obtained for 72 CD risk variants using odds ratios (OR) estimated from GWAS meta-analysis[13] (Supplementary methods). Smoking status was recorded as "Current", "Ex", or "Never" smoked with "Never" smoked as the reference. Smoking risk was incorporated into the REGENT analysis using ORs from a large case-control study[16]. FDRs in the highest and lowest quartiles of the risk score were invited to undergo VCE and provide stool and blood samples for biomarker analysis.

Video capsule endoscopy (VCE)

Capsule endoscopies were performed in the Endoscopy Unit at GSTT following written informed consent using MiroCam™ (Intromedic, Seoul, Korea) VCE system (Supplementary methods). Two validated scoring systems were used to quantify SI inflammation: the Lewis Score[18] and the Capsule Endoscopy Crohn's Disease Activity Index (CECDAI)[19]. SI transit time was defined as the passage from the first duodenal to the first caecal image, in minutes.

Biomarker assays

Stool samples were collected between 1-72 hours prior to the administration of laxatives for VCE and were stored at -80°C before faecal calprotectin (FC) extraction and ELISA analysis (Buhlmann EK-CAL Calprotectin, Buhlmann Laboratories, Schönenbuch, Switzerland). Approximately 20ml of blood in serum separator tubes (BD Vacutainer® SST™, BD Diagnostics, Oxford, UK) was collected from participants when they attended for VCE. Samples were centrifuged at room temperature for 10 minutes at 1300g. Sera were stored at -80°C prior to analysis for high-sensitivity C-reactive protein (hs-CRP), anti-saccharomyces cerevisiae antibodies (ASCA), and 18 additional cytokines, growth factors and cell adhesion molecules (Supplementary methods).

Statistical analysis

All statistical analyses were performed using R Project for Statistical Computing[20]. For the predictive modelling we used the R package caret[21], and its dependencies glmnet[22] and ranger[23].

Explanatory modelling: Logistic regression was used to model SI inflammation with dichotomized Lewis score from the VCE (<135 = normal/no inflammation, ≥135 = abnormal/intestinal inflammation). Explanatory variables of age, gender, genetic risk, smoking, CD family burden ("single" or "multiple" members affected with CD), VCE transit time and biomarkers were modelled

using stepwise selection applying Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), which seeks a more stringent model. Further details are given in **Supplementary methods**.

Predictive modelling: Machine learning was used to generate a predictive model to estimate the utility of the genetic, environmental and biomarker variables to identify individuals at high risk of SI inflammation. We divided the data into a training sample (2/3rds) for model building and a test sample (1/3rd) for evaluation of the model's predictive performance and compared two machine learning techniques, Elastic Net and Random Forest (Supplementary methods). For each approach, 20 repeats of 5-fold cross validation were used on the training data to optimise fit. The model was then applied to the test sample to attempt to classify FDRs with and without inflammation. Predictive performance was assessed with the area under the receiver-operator characteristic curve (AUC). Further details are given in **Supplementary methods**.

An updated SNP list: sensitivity test

To update our SNP list in line with more recent GWAS studies [24] we performed a sensitivity test to compare the effect of using a larger set of risk SNPs with updated risk score (see Supplementary methods).

Results

Calculating the high and low risk quartiles

The relative risk (RR) of CD in the 454 asymptomatic FDRs generated from 72 genetic risk markers and smoking status (**Figure 1**) ranged from 0.03–38.3 and was significantly higher than expected in

the general population (supplementary **Figure S1**) (p=0.013). Demographic information for each risk quartile is shown in supplementary **Table S2**. FDRs in the highest and lowest risk quartiles (n=228) were invited to participate in phase 2 of the study (**Figure 1**), of these 81 were unable to attend for varying reasons including travel and work commitments (**Figure S2**). The remaining 147 (64%) underwent VCE, 81 from the highest risk quartile and 66 from the lowest risk quartile (**Table 1**). For subsequent modelling, the risk score in the high and low risk quartiles was separated into a genetic score and smoking status.

Fecal and serum biomarker analysis

FC measurement was performed on 134 returned samples from FDRs in phase 2. Where adequate sample was provided, sera were analysed for 21 additional biomarkers resulting in complete biomarker data on 124 individuals (**Figure S2**). Six biomarkers had near zero variance among these individuals and were not analysed further (supplementary **Table S3**). A comparison of the demographics between those individuals with complete and incomplete data demonstrated no significant differences for age, gender or relationship to proband although some differences were observed for smoking status and CD family history (**Table 1**).

Video capsule endoscopy findings

All 147 FDRs in phase 2 underwent VCE and in 144 the caecum was reached (supplementary **Figure S2**). There was one capsule retention, managed conservatively with prokinetic agents. The most common abnormal finding was of small aphthous ulceration in the distal ileum. No strictures were identified. Marked inflammation typical of CD (>150 aphthous ulcers throughout the SI) was found in only one FDR, in the high-risk group. The majority (93%) of the lowest risk group had no SI inflammation (Lewis score <135) compared with 67% of the highest risk group (p=0.00037, **Table 1**).

In the high-risk group, 9% had moderate-severe inflammation (Lewis score ≥790) compared with none in the low risk group (p=0.016). Less than half of those with moderate-severe SI inflammation (45.6%) had a raised FC (>=50ug/g), which fell to 6% in those without SI inflammation (**Figure S3**).

All 11 participants with moderate-severe small bowel inflammation were offered follow-up in the IBD clinic at GSTT to discuss ongoing surveillance and the need for treatment. Of these, 8 participants attended for at least one visit. At the 3-year follow-up one of these had been diagnosed with and treated for CD. Five-year follow-up of the entire cohort will be completed in the next 12 months.

Characteristics of the highest and lowest risk quartiles

Of the 147 FDRs who underwent VCE, 124 had complete data for all biomarkers and were included in all analyses (supplementary **Table S4**). Among these variables, smoking and Lewis score showed a significant difference by risk quartile (**Table 1**). The Lewis and CECDAI capsule scores were highly correlated (Pearson's r=0.89, 95% CI=0.85-0.92, p<0.01) and when each measure was dichotomized (Normal, Abnormal) only 3 samples were classified differently. For all modelling, we used the dichotomized Lewis score as our outcome, as this has previously been shown to correlate well with FC whereas the CECDAI did not[25].

Explanatory modelling

SI inflammation (based on Lewis score) was modelled using all genetic, environmental and biomarker variables to explain the observed VCE data. Using stepwise selection and AIC we reduced the number of predictor variables to a best explanatory subset of 14 (R2=0.72). Stepwise selection using the more stringent BIC yielded 3 variables (R2=0.47) including FC, CD family history, and

genetic risk score (supplementary **Table S5**). Although regression modelling can quantify the variance in the outcome explained by the predictors, these classical statistical approaches do not account for correlation among predictors (**Figure S4**) and provide no indication of the predictive ability of the derived model.

Predictive modelling

Predictive models using elastic net and random forest were built on a training sample of 83 FDRs, then tested on the remaining sample of 41 FDRs. The predictive performance of both methods was assessed with the area under the receiver-operator characteristic (ROC) curve (**Figure 2**). The random forest (AUC=0.87, 95% CI=0.75 – 1.00; Accuracy=0.73, 95% CI=0.57 – 0.86) performed slightly better than the elastic net (AUC=0.80, 95%CI=0.62 – 0.98; Accuracy=0.68, 95% CI=0.52 – 0.82), correctly classifying an additional 2 test samples (1 normal/un-inflamed, 1 abnormal/inflamed) among the 41 unseen test samples (**Table 2**, **Table S6**).

Genetic risk score improves predictive performance

The elastic net model reduced the full set of predictor variables to the three most important variables: CD family history (OR=1.31), genetic risk score (OR=1.14), and FC (OR=1.04) (supplementary **Table S7**). Although it is not possible to assign p-values or confidence intervals to individual predictors in a cross-validated elastic net solution[26], interestingly, these were the same predictors selected by stepwise BIC in the explanatory modelling (**Table 3**, supplementary **Table S5**). Genetic risk score was the second most important variable in the random forest model. Although it is possible to assign p-values to importance in a random forest, in order evaluate the contribution of genetic risk to both models in a comparable way, we built the two models with and without genetic risk score. In both models, excluding genetic risk score reduced predictive performance for SI

inflammation. An elastic net excluding genetic risk score yielded a two-variable model: FC (OR=1.09) and CD family history (OR=1.63) with significantly lower AUC (0.78, 95% CI: 0.55 – 1.00) compared to the original full elastic net model (AUC=0.80, 95% CI=0.62 – 0.98 (Mann-Whitney U test derived[27])). Likewise, performance of a random forest model excluding genetic risk score was significantly lower (AUC=0.83, 95% CI: 0.67 – 0.99) than the original random forest model (AUC=0.87, 95% CI=0.75 – 1.00).

Updated SNP set and risk scores

Using updated SNP risk data from Jostins *et al* 2012 we found that the original and updated risk scores and the updated and expanded risk scores were highly correlated (r = 0.952, **Figure 3**, *r* = 0.72 **Figure S5** respectively). Also, the risk of being in the highest risk quartile, given the presence of SI inflammation, was very similar under the original SNP set and risk scores OR = 6.48 (95% CI: 2.08 - 20.19), and the expanded SNP set and risk scores OR = 6.21 (95% CI: 1.28 - 30.18) (See **Supplementary material** for further details).

Discussion

We found that the combination of CD family history, genetic risk, and level of FC was a good predictor of SI inflammation in our cohort of FDRs. Both machine learning classifiers performed similarly, with significant prediction of SI inflammation, and both put CD family history, genetic risk and FC as top predictors, which increases confidence in the model's utility. CD family history was the strongest predictor in the elastic net model and ranked 5th most important predictor in the random forest. By design, all our study participants had at least one relative with CD; we found that a stronger family history (≥2 relatives with CD) increased the risk for SI inflammation, agreeing with

previous studies[4]. However, the number of affected family members is a crude measure of familial risk. Ideally, total family size, structure, and age of affected individuals should also be incorporated[28].

We found that genetic risk score was a significant predictor of SI inflammation. The predictive accuracy of our elastic net classifiers with or without genotype were both within the 0.7–0.8 range generally considered to be acceptable discrimination[29]. Performance of the random forest models with or with genotype were both in the 0.8-0.9 range considered excellent discrimination. On inclusion of genetic risk in the model elastic net estimates showed a 2% and random forest a 4% increase in AUC, thus predicting an additional 2 and 4 cases respectively in every 100 randomly selected pairs. In computing the genetic risk score we used 72 CD-associated genetic variants known at the time of clinical assessment[13], and the single best-understood lifestyle factor, smoking. Since then the number of genetic loci associated with IBD has risen to 240 [24,30,31]. To estimate the effect of updating risk variants, we used the set of CD associated SNPs identified in the Jostins *et al* (2012) GWAS meta-analysis for which we had available data, given that we had used the same genotyping array (Immunochip). We showed a high correlation (r2=0.952), between our original risk scores used to select FDRs for VCE and the updated risk scores based on this expanded more recent set of risk SNPs. Updating the genetic risk panel as more risk loci are discovered could increase the performance of the model further.

We found that 35% of FDRs had an elevated FC (≥50 µg/g), which has previously been observed in asymptomatic FDRs[12]. As FC is part of the diagnostic work-up for CD[10], and increased levels predict relapse[32], our findings of its predictive utility for SI inflammation suggest it is a promising biomarker for those at greatest risk for CD. However, we observed that a cut-off of ≥50 µg/g would include many false positives, whereas a threshold of >250 µg/g, which has been suggested as appropriate for screening asymptomatic individuals,[33] resulted in a high false negative rate in our FDR cohort

(>80%, **Figure S3**). Therefore, our data demonstrates that FC alone has limited use in classifying asymptomatic individuals at greatest risk for developing CD, whereas our predictive model benefits from the full range of information contained in the biomarker in the context of all other predictors.

Mild SI inflammation found at VCE has previously been reported at a rate of 24% of asymptomatic FDRs, but the subsequent development of CD was not determined[34]. A study of 38 FDRs who underwent ileocolonoscopy found mild endoscopic and histological inflammation in 26% and CD in 13%. Those with mild inflammation underwent repeat ileocolonoscopy after a mean of 53 months without endoscopic or histological progression of inflammation[10]. In our study 26/124 (21%) FDRs had abnormal Lewis scores, which is broadly consistent with these data. However, it is not certain whether these features are predictive of future development of CD. Long term follow-up of our FDR cohort may provide further information regarding risk of developing overt CD as opposed to asymptomatic SI inflammation.

Although we used an unseen subset of samples to test the predictive performance of our model, this test sample was subject to the same design decisions as the training samples, as well as any study-specific idiosyncrasies. Ultimately, our finding will require external validation by replication in an independent sample.

A future replication study could benefit from the inclusion of an updated list of IBD-associated SNPs (most IBD loci confer risk for both CD and UC[31]), and a more comprehensive picture of the environmental risk (i.e., in addition to smoking, medication, diet, stress, sleep, physical activity[9]). Finally, this study did not assess the gut microbiome, a likely predictor of risk for CD and potential target for intervention. The Genetic, Environmental and Microbiome (GEM) Project, currently

recruiting 5000 CD FDRs internationally, will hopefully bring further insights into pre-clinical CD with the combination of all of these factors[35].

In parallel with our increasing understanding of the specific genetic and environmental risk factors that combine to make an individual's immune system hostile to commensal gut flora, a clinically useful tool for detecting those at greatest risk for CD before the presentation of overt symptoms would prioritise patient screening and follow up. Early detection opens up the possibility for early intervention, additional targets for drug development, and disease prevention. Our study suggests that a CD prediction tool can be built from a small set of biomarkers, known genetic risk variants, and family history of CD. Inclusion of risk factors from the most recent findings will improve prediction accuracy.

References

1. Halme L, Paavola-Sakki P, Turunen U, et al. Family and twin studies in inflammatory bowel disease. World J Gastroenterol 2006;12:3668-3672.
2. De Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet 2017;49.
3. Gordon H, Trier Moller F, Andersen V, et al. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. Inflamm Bowel Dis 2015;21:1428-1434.
4. Kevans D, Silverberg MS, Borowski K, et al. IBD Genetic Risk Profile in Healthy First-Degree Relatives of Crohn's Disease Patients. J Crohns Colitis 2016;10:209-215.
5. Molodecky NA, Soon IS, Rabi DM, et al. Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. Gastroenterology 2012;142:46-54.e42.
6. Ananthakrishnan AN. Epidemiology and risk factors for IBD. Nat Rev Gastroenterol Hepatol 2015;12:205-217.
7. Baumgart DC, Sandborn WJ. Crohn's disease. Lancet 2012;380:1590-1605.
8. Mahid SS, Minor KS, Soto RE, et al. Smoking and Inflammatory Bowel Disease: A Meta-analysis. Mayo Clin Proc 2006;81:1462-1471.
9. Hedin CR, Stagg AJ, Whelan K, et al. Family studies in Crohn's disease: new horizons in understanding disease pathogenesis, risk and prevention: Figure 1. Gut 2012;61:311-318.
10. Sorrentino D, Avellini C, Geraci M, et al. Tissue Studies in Screened First-degree Relatives Reveal a Distinct Crohn's Disease Phenotype. Inflamm Bowel Dis 2014;20:1.

11. Biancone L, Calabrese E, Petruzziello C, et al. A family study of asymptomatic small bowel Crohn's disease. Dig Liver Dis 2014;46:276-278.

12. Teshima CW, El-Kalla M, Turk SA, et al. 220 Asymptomatic First Degree Relatives of Crohn's Patients Display Endoscopic Small Intestinal Lesions Independent of Their Gut Permeability Status. Gastroenterology 2012;142:S-54.

13. Franke A, McGovern DPB, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118-1125.

14. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther 2010;13:101.

15. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 2007;81:559-575.

16. Brant SR, Wang M-H, Rawsthorne P, et al. A Population-Based Case-Control Study of CARD15 and Other Risk Factors in Crohn's Disease and Ulcerative Colitis. Am J Gastroenterol 2007;102:313-323.

17. Crouch DJ, Goddard GH, Lewis CM. REGENT: a risk assessment and classification algorithm for genetic and environmental factors. Eur J Hum Genet 2013;21:109-111.

18. GRALNEK IM, DEFRANCHIS R, SEIDMAN E, et al. Development of a capsule endoscopy scoring index for small bowel mucosal inflammatory change. Aliment Pharmacol Ther 2007;27:146-154.

19. Gal E, Geller A, Fraser G, et al. Assessment and Validation of the New Capsule Endoscopy Crohn's Disease Activity Index (CECDAI). Dig Dis Sci 2008;53:1933-1937.

20. R: The R Project for Statistical Computing. https://www.r-project.org/.

21. CRAN - Package caret. https://cran.r-project.org/web/packages/caret/index.html.

22. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1-22.

23. Wright MN, Ziegler A. **ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. J Stat Softw 2017;77:1-17.

24. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491.

25. Koulaouzidis A, Douglas S, Plevris JN. Lewis Score Correlates More Closely with Fecal Calprotectin Than Capsule Endoscopy Crohn's Disease Activity Index. Dig Dis Sci 2012;57:987-993.

26. Wu TT, Chen YF, Hastie T, et al. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 2009;25:714-721.

27. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q J R Meteorol Soc 2002;128:2145-2166.

28. Yasui Y, Newcomb PA, Trentham-Dietz A, et al. Familial Relative Risk Estimates for Use in Epidemiologic Analyses. Am J Epidemiol 2006;164:697-705.

29. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. https://www.wiley.com/en-gb/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473. Accessed December 29, 2018.

30. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet July 2015.

31. de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet 2017;49:256-261.

32. Ikhtaire S, Shajib MS, Reinisch W, et al. Fecal calprotectin: its scope and utility in the management of inflammatory bowel disease. J Gastroenterol 2016;51:434-446.

33. D'Haens G, Ferrante M, Vermeire S, et al. Fecal calprotectin is a surrogate marker for endoscopic lesions in inflammatory bowel disease. Inflamm Bowel Dis 2012;18:2218-2224.

34. Teshima CW, Goodman KJ, El-Kalla M, et al. Increased Intestinal Permeability in Relatives of Patients With Crohn's Disease Is Not Associated With Small Bowel Ulcerations. Clin Gastroenterol Hepatol 2017;15:1413-1418.e1.

35. GEM Project | The Crohn's and Colitis Canada GEM Project. http://www.gemproject.ca/.

Table 1. Distribution of demographics, relationship and Lewis score by risk quartile

Table 2. Classification tables: cross-tabulations of observed (reference) and predicted outcome.

Table 3. Relative predictor importance.

Figure legends

Figure 1. Difference in relative risk between the quartiles (see Methods for explanation of calculation). n = 114 first quartile (lowest risk), 113 second quartile, 113 third quartile, 114 fourth quartile (highest risk).

Figure 2. Classifier evaluation by ROC for test sample. Test sample performance was similar for both models: elastic net sensitivity=0.75, specificity=0.67; random forest sensitivity=0.88, specificity=0.70. ROC curve: True positive rate (or Sensitivity) = TP / (TP+FN); True negative rate (or 1-Specificity) = 1 - (TN / (TN+FP)).

Figure 3. Correlation between the original risk score - which incorporates 69 SNPs and corresponding risks from Franke *et al*, 3 x *NOD2* risk variants (Materials and Methods), and smoking - and the updated Jostins et al (2012) risk scores for the same set of SNPs/variants.

Table 1. Distribution of demographics, relationship and Lewis score by risk quartile

	HIGHEST RISK QUARTILE			LOWEST RISK QUARTILE			
	Incomplete data	Complete data	Difference	Incomplete data	Complete data	Difference	Difference between highest and lowest risk quartiles with complete data
Number	47	67		57	57		n/a
Relationship to proband			p = 0.03			p = 0.12	p = 0.90
sibling	22 (47%)	26 (39%)		25 (43%)	23 (40%)		
offspring	24 (51%)	29 (43%)		29 (50%)	22 (39%)		
parent	1 (2%)	12 (18%)		4 (7%)	11 (19%)		
Gender			p = 0.39			p = 0.44	p = 0.94
female	27 (57%)	45 (67%)		32 (56%)	37 (65%)		
male	20 (43%)	22 (33%)		25 (43%)	20 (35%)		
Age			p = 0.18			p = 0.18	p = 0.94
mean	34.8	37.4		34.7	37.5		
median	*34.3*	*37.4*		*33.2*	*36.2*		
Smoking status			*p = 7.0×10^{-4}			p = 0.70	*p = 1.1×10^{-6}
current	29 (62%)	23 (34%)		3 (5%)	6 (11%)		
ex	16 (34%)	25 (37%)		8 (14%)	8 (14%)		
never	2 (4%)	19 (28%)		46 (80%)	43 (75%)		
CD family history			*p = 2.2×10^{-9}			*p = 2.8×10^{-3}	p = 0.78
single	15 (32%)	59 (88%)		38 (67%)	52 (91%)		
multiple	32 (68%)	8 (12%)		19 (33%)	5 (9%)		
Lewis score			n/a			n/a	*p = 3.7×10^{-4}
Normal (<135)	n/a	45 (67%)		n/a	53 (93%)		
Abnormal (≥135)	n/a	22 (33%)		n/a	4 (7%)		

Incomplete data = subset of the respective risk quartile with missing VCE or biomarker data; Complete data = subset of the respective risk quartile with complete data on all measures; Difference = test of significant differences within (subsample with incomplete data vs subsample with complete data) and between (subsample with complete data: highest vs lowest) risk quartiles. Percentages are column percentages within each subheading. p-values relate to a test for equality of proportions, a *t*-test of mean differences, a Pearson's chi-squared test, or a Fisher's exact test as appropriate. * = significant after Bonferroni multiple test correction (0.05 / 16)

Table 2. Classification tables: cross-tabulations of observed (reference) and predicted outcome

Elastic net

		Reference	
		Normal	Abnormal
Prediction	Normal	22	2
	Abnormal	11	6

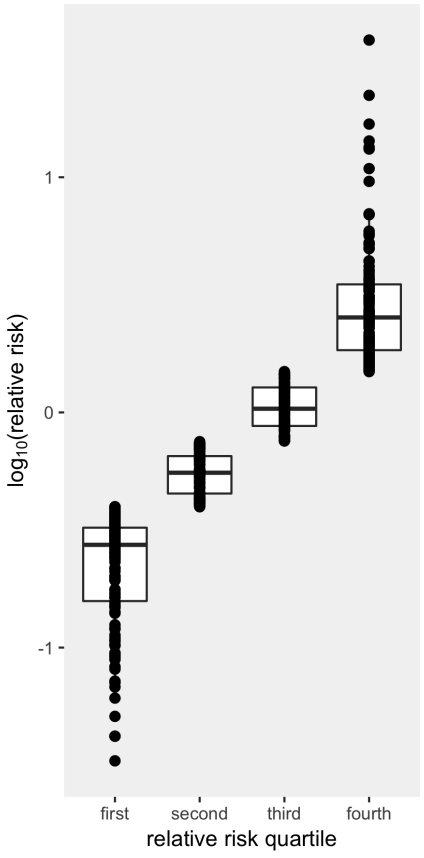
Random forest

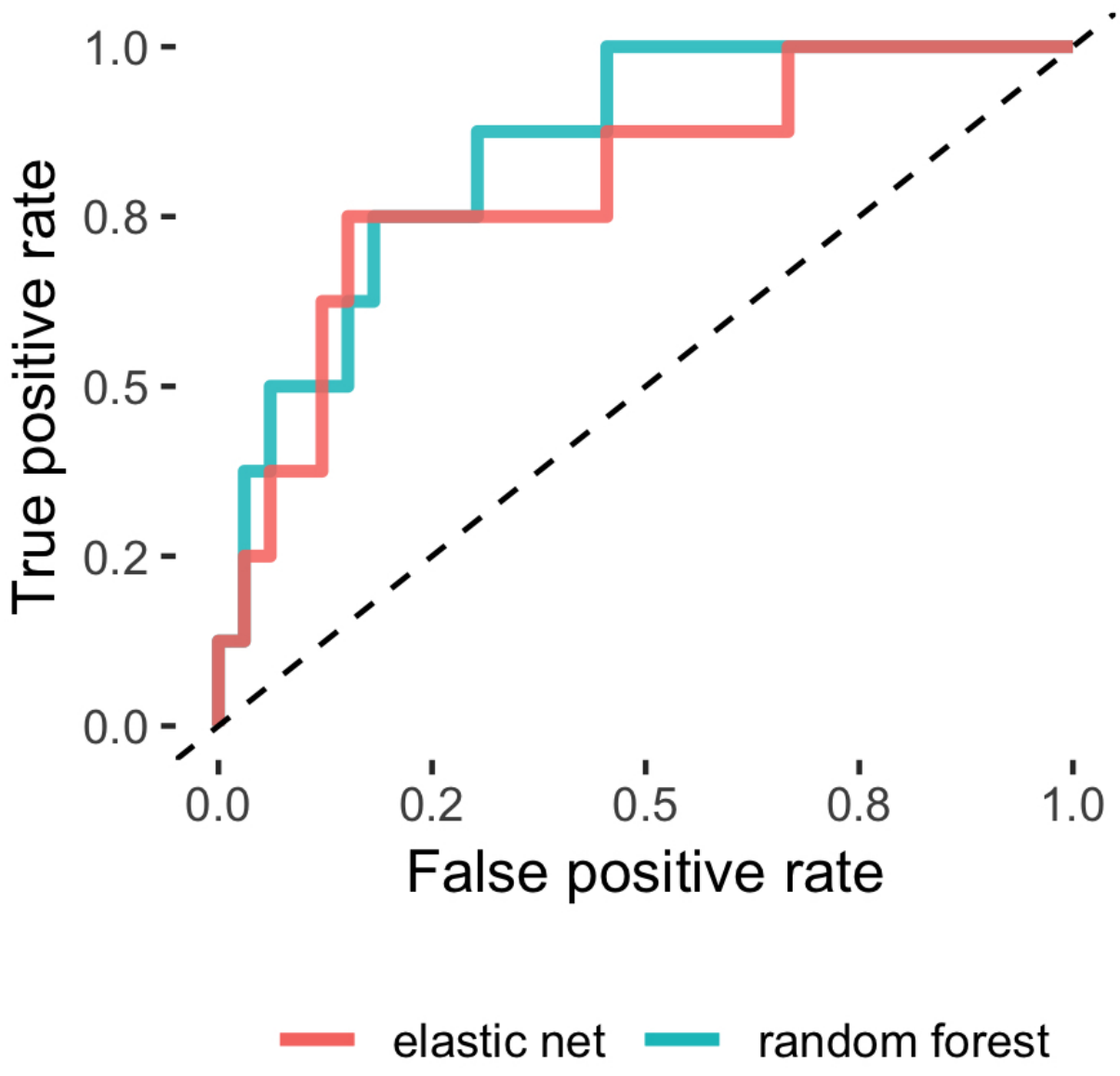
		Reference	
		Normal	Abnormal
Prediction	Normal	23	1
	Abnormal	10	7

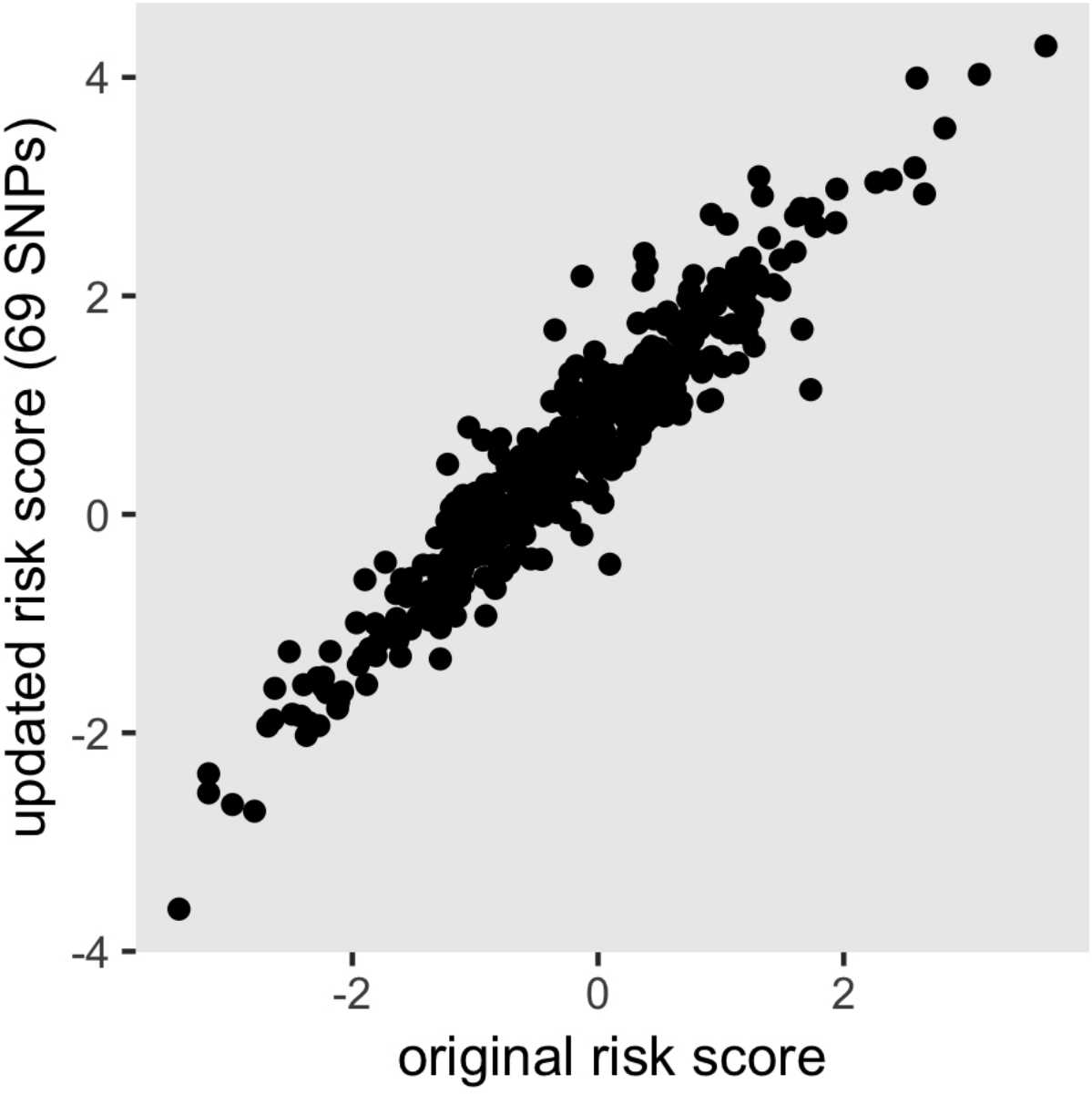
Table 3. Relative predictor importance

	Elastic net		Random forest	
Rank	*Predictor*	*Importance*	*Predictor*	*Importance*
1	CD family history	0.27	faecal calprotectin	0.0967
2	genetic risk score	0.13	genetic risk score	0.0264
3	faecal calprotectin	0.04	hs-CRP	0.0151
4			IL6	0.0043
5			CD family history	0.0023

Note: Only top 5 predictors shown for the random forest. Importance is an indicator of a particular variable's contribution to model performance. The random forest importance is the mean decrease in accuracy given by the difference in error rate after permuting the particular variable, averaged over all trees. The elastic net importance is the absolute value of the regression coefficient *t*-statistic.




1.0
0.8
0.5
0.2
0.0
True positive rate
0.0
0.2
0.5
0.8
1.0
False positive rate
elastic net
random forest



SUPPLEMENTARY MATERIAL:

Predictors of elevated gut inflammation in asymptomatic first-degree relatives of patients with Crohn's disease

*Kirstin M. Taylor, *Ken B. Hanscombe, *Natalie J. Prescott, Raquel Iniesta, Matthew Traylor, Nicola S. Taylor, Steven Fong, Nicholas Powell, Peter M. Irving, Simon H. Anderson, Christopher G. Mathew, Cathryn M. Lewis, Jeremy D. Sanderson

* Joint first authors

1. Supplementary Materials and Methods

SNP genotyping

DNA samples were genotyped on the Immunochip, a custom Illumina Infinium array containing 196,524 SNPs and small insertion/deletions selected mainly from GWAS analysis of 12 immune-mediated diseases[1]. The immunochip included 70 out of the 71 known significant CD SNPs based on genome wide association studies (GWAS) at the time of study design (rs736289 was not present on Immunochip)[2]. The Immunochip also included the three major CD-risk variants in nucleotide oligomerisation domain 2 gene (NOD2) rs2066844 (p.R702W), rs2066845 (p.G908R) and rs2066847 (p.L1007insC) and these were used in preference to the NOD2 tagging SNP (rs2076756) from Franke et al due to their well-established role in disease, bringing the total number of SNPs to 72. The three NOD2 risk variants were combined to model the inflated risk of NOD2 risk-variant homozygotes and compound heterozygotes as previously described[3].

SNPs with >3% missing genotype data and Hardy-Weinberg equilibrium outliers ($p \leq 1 \times 10\text{-}4$) were excluded. Individuals missing >3% of SNPs or with gender inconsistency were excluded. Population structure was assessed by principal components analysis in PLINK as previously described[4], derived from HapMap 3 CEU population - Utah residents with Northern and Western European ancestry from the CEPH (Centre d'Etude du Polymophisme Humain). No outliers were identified.

In total, 454 FDRs were successfully genotyped (61% female; median age 34, range 18-55; 40% were siblings, 46% offspring, 14% parents of probands; 44% were current or ex-smokers).

Calculation of risk

Risk modeling was performed within the R package, REGENT, incorporating published gene-environment risk factor and disease statistics to categorize risk using a confidence interval (CI)-

based approach within a simulated population[5,6]. Genetic risk factors (allelic odds ratios (ORs), allele frequencies and sample sizes) were determined from the first genome-wide meta-analysis in Crohn's disease[7] (See supplementary **Table S1**). The risk for *NOD2* variants was more complicated due to the enhanced risk of homozygotes and compound heterozygotes. Extensive case-control data for *NOD2* from our previously published studies[8,9] was modeled using logistic regression in R to assess the model that provided the best fit in REGENT – simply counting the numbers of mutations proved best. Environmental risk factors required ORs, standard errors and the proportion of the population exposed to the risk factor, based on a large case-control study[3]. The first stage of REGENT involved simulation of a population-distribution of disease risk using a multiplicative model. Confidence intervals were used to classify the risk profiles into four categories: reduced, average, elevated and high-risk. The second stage of REGENT applied this simulated model to individual level data. This gave a disease risk in relation to the general population.

An updated SNP list: Sensitivity testing

The Immunochip employed here for genotyping was also used for a major IBD GWAS meta-analysis that was published after the commencement of our study (Jostins et al, 201224). This described a total of 163 loci for IBD, which included nearly 70 new loci for CD (total 140 CD loci), all with updated risks based on a larger patient population. To investigate how these updated risk values affected our original model we performed a sensitivity test. We compared the distribution of normal versus abnormal inflammation across highest and lowest risk quartiles for our original risk score based on 69 SNPs using effect sizes from Franke et al 2010, (including NOD2 and smoking risk) with an updated risk score including the same 69 SNPs but using the effect sizes from Jostins et al., 2012. Then we compared the updated risk score to an expanded risk score including 134 out of 140 CD associated SNPs that passed QC in our data using effect sizes from Jostins et al., 2012. Where a GWAS identified variant was not available on the Immunochip, we used the reported Immunochip tag SNP24. For each individual, a genetic risk score was calculated as the weighted sum of the number of risk alleles using the log of the odds ratios. For tagging SNPs, the log(OR) was scaled down by the linkage disequilibrium between the GWAS-identified SNP and the Immunochip tag. As previously, we included the combined (homozygous and compound heterozygous) risk of the NOD2 variants and smoking in the updated and expanded risk SNP scores.

We found that the original and updated risk scores and the updated and expanded risk scores were highly correlated (r = 0.952, **Figure 3**, *r* = 0.72 **Figure S5** respectively). Also, whilst we were unable to go back and re-select our highest and lowest risk quartiles from the FDR group for VCE, we calculated that the proportion of individuals in the highest risk quartile that had SI inflammation on VCE were comparable when using the original risk score (33%), the updated risk score (33%), and the expanded risk score (29%). The risk of being in the highest quartile, given the presence of SI inflammation, was found to be very similar under the two different risk models: the original risk score OR = 6.48 (95% CI: 2.08 - 20.19), the expanded risk score OR = 6.21 (95% CI: 1.28 - 30.18).

Biomarkers

Approximately 20ml of blood in serum separator tubes (BD Vacutainer® SST™, BD Diagnostics, Oxford, UK) was collected from participants when they attended for capsule endoscopy. Samples were spun at room temperature for 10 minutes at 1300 relative centrifugal force (2500 rotations per minute) as per the manufacturer's instructions. Aliquots of 1-2ml of serum were frozen at -80°C. Serum was analysed for high-sensitivty C-reactive protein (hs-CRP) using the CardioPhase assay, Siemen's Healthcare Diagnostics, Erlangen, Germany). Serum anti-saccharomyces cerevisiae antibodies (ASCA) were measured using the QUANTA Lite ASCA IgA and IgG ELISA assays (Inova Diagnostics, San Diego, California, USA). Serum cytokines and growth factors were analysed using the Randox Cytokine Chip Array Custom X, and adhesion molecules using the Randox Adhesion Molecules Array (both Randox Laboratories, Crumlin, Co Antrim, UK). Stool samples were collected before the administration of laxatives for VCE and were frozen at -80°C before faecal calprotectin (FC) extraction and ELISA analysis (Buhlmann EK-CAL Calprotectin, Buhlmann Laboratories, Schönenbuch, Switzerland). Adequate faecal specimens for calprotectin measurement were returned from 134 FDRs. Mean FC levels were similar in both risk groups (86 µg/g in high vs 90 µg/g in low risk group, p=0.92) but were higher in those with moderate-severe small intestinal inflammation (186 µg/g) vs normal-mild inflammation (33 µg/g) p=0.03.

Video capsule endoscopy

The capsule endoscopies were performed in the Endoscopy Department of St Thomas' Hospital following written informed consent. The MiroCam™ (Intromedic, Seoul, Korea) system was used for all capsule endoscopies. The MiroCam™ capsule (measuring 11 mm × 24 mm) was swallowed with water containing simethicone 40mg. Study participants were asked to avoid non-steroidal

anti-inflammatory drugs (NSAIDs) in the 3 months prior to capsule endoscopy. To improve diagnostic yield, two days before capsule endoscopy, FDRs began a low fibre diet, and the following day they commenced a clear liquid diet, followed by an overnight fast. A laxative, Picolax (Ferring, West Drayton, UK), containing sodium picosulphate and magnesium citrate, was taken in two doses the day before the procedure. Images were analyzed using a dual reading frame at a maximum speed of 24 frames per second. Each recording was read by two independent observers, fully trained in VCE, and blinded to each other and the risk status of the subject. Where there was a discrepancy in the findings, a third observer adjudicated.

2. Explanatory modelling

After removal of near-zero variance predictors (ASCA IgG, ASCA IgA, IL10, IL1a, IL4, INFy), we performed stepwise selection using the R package MASS with step direction = "both". Near zero variance evaluation is described below. We first fitted multivariate logistic regression models with dichotomized Lewis score (<135 = Normal i.e. without intestinal inflammation, ≥135 = Abnormal, i.e. with intestinal inflammation) as response variable, and age, sex, genetic risk score, smoking status, CD family history, VCE capsule transit time and the biomarkers as explanatory variables. FDRs by definition have one CD-affected relative. However, some participants had more than one affected family member. We included this additional family burden in our predictive modelling. Therefore, we defined the extent of CD family history as having “single” or “multiple” members affected with CD. Applying generalised linear regression models to a dataset with a large number of predictor variables relative to the number of samples is not optimal for several reasons including over-fitting (explaining noise as well as signal in the sample), multi-collinearity (correlation and potential redundancy among predictors), and low interpretability28. We used stepwise model selection to address this overfitting, multicollinearity, and model complexity. At each step in the iterative selection procedure, a variable is considered for addition to (or subtraction from) the current set of explanatory variables based on a model comparison criterion. We compared two criteria: Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). However, this approach does not eliminate the problems of overfitting, multicollinearity, and model complexity, especially with a small sample and large number of measured variables. In addition, explanatory models can quantify the relationship between predictors and outcome in the particular sample collected (e.g. regression coefficients, and total variance explained), but they give no measure of the predictive performance of the putative

predictors in unseen cases. Given that the ultimate aim of the study was to estimate the utility of the genetic, environmental and biomarker variables to identify high risk individuals, we went on to derive a series of predictive models. Machine learning is specifically designed to deal with the large number of variables relative to sample size problem, and includes techniques to address the low events-per-variable ratio (small number of cases). Supplementary **Table S5** shows the results of both Akaike information criterion (AIC) and the Bayesian information criterion (BIC) best explanatory selection.

3. Predictive modelling

Our goal was to develop a predictive model for SI inflammation, achieving a balance between interpretability (by reducing the number of predictor variables) and predictive ability. Thus we used machine learning, which has techniques for variable subset selection and estimation of how well a given model will perform at predicting future data29. Machine learning finds structure in data and addresses over-fitting when there are a large number of predictors. More generally, it is a set of techniques that improve performance on a specified task (e.g. classifying absence/presence of inflammation) with experience (exposure to data). We divided the data into a training sample (2/3) for model building and a test sample (1/3) for evaluation of model predictive performance and compared two machine learning techniques:

Elastic net. The elastic net is an extension of the basic regression framework that allows selection of the most important subset of predictors. It mixes a ridge penalty (which shrinks the coefficients of correlated predictors towards each other) and a LASSO penalty (which selects one among a group of correlated predictors and shrinks the coefficients of the others to zero) to perform variable selection28. We used 20 repeats of 5-fold cross validation to estimate model parameters (penalty mixing factor (α) and penalty strength (λ)) that optimised the model's prediction performance, i.e., its ability to correctly classify individuals with and without inflammation. We measured prediction performance with the area under that receiver-operator characteristic curve (AUC: a plot of the true positive rate against the false positive rate for different cut-offs of the model estimated probability of being Abnormal) and accuracy (the fraction of test sample predictions that are true). The absolute value of the t-statistic for each parameter in the model is used to judge relative variable importance.

Random forest. A random forest is a collection of classification or regression trees. A tree is a series of splitting rules. At each split in a particular tree, from a random subset of predictors a single predictor is chosen that produces branches with the best split of Normal and Abnormal samples. Each resulting branch is then split until it ends in a "leaf" containing only (or mainly) Normal or Abnormal training samples. A test sample is then pass through each tree, and is assigned the (majority) class of the leaf on which it lands. Every tree in the forest produces a prediction and these predictions are combined to give a single consensus prediction for the individual28 29. We used 20 repeats of 5-fold cross validation on the training data to select the optimum number of randomly selected predictors (mtry) that maximized the AUC. Overall variable importance was determined by permuting predictors one at a time and measuring the mean decrease in accuracy averaged over all trees.

Machine learning considerations

Cross-validation

In *n*-fold cross validation, the sample is randomly partitioned into 5 subsamples. For i = {1, 2, …, *n*}, the *i*th subsample is left out and the model is fitted (or estimated) on the *n-1* remaining subsamples. The derived model is then evaluated on the *i*th left out subsample. The model evaluation metric (e.g. AUC) is then averaged across all *i* iterations. This was repeated over a range of values of α and λ in order to find the combination of penalty mixing and strength that produces the best elastic net predictive performance. Similarly, for the random forest the cross-validation was repeated over a range of values for *mtry* to find the optimal number of randomly selected variables to search among at each branching node. As the folds are randomly selected by design, we repeated the entire procedure 20 times to achieve a stable solution for both predictive models.

Near zero variance predictors

If most individuals in a dataset have a single unique value for a particular variable (i.e., a "near-zero variance predictor"), including this variable in regression modelling can cause the model to become unstable or have undue influence on the model (see supplementary materials for more detail). If all individuals have the same value, the variable has zero variance and provides no useful information for model building. Near-zero variance is judged on two criteria: few unique values relative to the number of samples, and the ratio of the count of the most common value to the

count of the second most frequent value. We used caret::nearZeroVar options freqCut = 90/10 for ratio of most common variable value to second most common variable value and uniqueCut = 20 for percentage of distinct values, to determine near-zero variance predictors. Effectively, a variable was dropped if its mode was 9 times more common than the next most frequent value and if less than 20% of the sample had a unique value.

Importance

Variable importance ranks the contribution of variables to a predictive model. The variable importance measure used for the elastic net is the absolute value of the *t*-statistic for each predictor in the underlying generalized linear model run on the training data. For the random forest, a predictor variable's importance is determined by the decrease in out-of-bag accuracy (averaged over all trees) after permuting the variable. The out-of-bag observations are the observations in the cross-validation fold used to evaluate model performance, during training. We used ranger::ranger option importance = "permutation" to calculate random forest predictor variable importance by permutation.

Test-train sample split

After performing data cleaning, including checks for near zero variance within and multicollinearity between predictors, we partitioned the data into training and test samples. We used 2/3 of participants in the training sample to tune model parameters and measure variable importance. In the remaining 1/3 we evaluated the predictive performance of each model. Using caret::createDataPartition, we preserved the Normal-Abnormal ratio (class distribution) in the training and test samples. We pre-processed the training data (centred and scaled) and applied the training data location and scale to the test data before evaluation.

Class imbalance

For the elastic net, imbalance in the training sample was addressed using a weighting: (Normal, Abnormal) → (Normal, Abnormal * (Normal count/ Abnormal count)). This is a standard approach to class imbalance in logistic regression. For the random forest, we used the caret::trainControl option sampling = "down" to down-sample the majority class (Normal) to better match the rarer class (Abnormals). As the random forest is an ensemble method (results are combined across trees to generate a single prediction), down-sampling does not result in loss of information. If sufficient

trees are generated, all samples will be included. We generated a random forest with 2000 classification trees.

References

1. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther 2010;13:101.
2. Franke A, McGovern DPB, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42.
3. Brant SR, Wang M-H, Rawsthorne P, et al. A Population-Based Case-Control Study of CARD15 and Other Risk Factors in Crohn's Disease and Ulcerative Colitis. Am J Gastroenterol 2007;102:313-323.
4. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 2007;81:559-575.
5. Crouch DJ, Goddard GH, Lewis CM. REGENT: a risk assessment and classification algorithm for genetic and environmental factors. Eur J Hum Genet 2013;21:109-111.
6. Goddard GHM, Lewis CM. Risk categorization for complex disorders according to genotype relative risk and precision in parameter estimates. Genet Epidemiol 2010;34:624-632.
7. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118-1125.
8. Prescott NJ, Fisher SA, Franke A, et al. A Nonsynonymous SNP in ATG16L1 Predisposes to Ileal Crohn's Disease and Is Independent of CARD15 and IBD5. Gastroenterology 2007;132:1665-1671.
9. Onnie CM, Fisher SA, Prescott NJ, et al. Diverse effects of the CARD15 and IBD5 loci on clinical phenotype in 630 patients with Crohn's disease. Eur J Gastroenterol Hepatol 2008;20:37-45.
10. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491.

Table S1. Genetic risk input file for REGENT

No	SNP	MAF	RR_het	RR_hom	Gene	Ncase	Ncontrol
1	imm_1_7801650	0.19	1.05	1.1	VAMP3	6333	15056
2	imm_1_67478546	0.068	0.38	0.14	IL23R	6333	15056
3	imm_1_114179091	0.093	0.79	0.63	PTPN22	6333	15056
4	imm_1_153496755	0.25	1.13	1.28	SCAMP3, MUC1	6333	15056
5	rs4656940	0.199	0.87	0.76	CD244, ITLN1	6333	15056
6	imm_1_171120083	0.246	1.22	1.49	TNFSF18	6333	15056
7	imm_1_195994265	0.302	1.04	1.08	DENND1B	6333	15056
8	imm_1_199144185	0.274	0.88	0.77	C1orf106	6333	15056
9	imm_1_205006527	0.157	1.12	1.25	IL10, IL19	6333	15056
10	1kg_2_25345971	0.326	1.06	1.12	DNMT3A	6333	15056
11	rs780094	0.418	1.15	1.32	GCKR	6333	15056
12	1kg_2_43660422	0.129	1.14	1.3	THADA	6333	15056
13	imm_2_61077763	0.42	1.14	1.3	C2orf74, REL	6333	15056
14	imm_2_102420881	0.231	1.19	1.42	IL18RAP	6333	15056
15	rs6738825	0.473	1.06	1.12	PLCL1	6333	15056
16	imm_2_230825118	0.187	1.12	1.25	SP140	6333	15056
17	imm_2_233849156	0.471	0.75	0.56	ATG16L1	6333	15056
18	rs13073817	0.322	1.08	1.17	0	6333	15056
19	imm_3_49696536	0.297	1.22	1.49	MST1	6333	15056
20	imm_5_40446341	0.394	0.75	0.57	PTGER4	6333	15056
21	1kg_5_72586890	0.4	0.89	0.8	0	6333	15056
22	imm_5_96270394	0.409	1.05	1.1	ERAP2, LRAP	6333	15056
23	imm_5_131812292	0.422	1.23	1.51	SLC22A4	6333	15056
24	imm_5_141459249	0.204	0.94	0.89	NDFIP1	6333	15056
25	imm_5_150250613	0.088	1.37	1.88	IRGM	6333	15056
26	imm_5_158719963	0.332	1.18	1.39	IL12B	6333	15056
27	1kg_5_173212448	0.429	0.93	0.86	CPEB4	6333	15056
28	rs17309827	0.361	0.91	0.83	0	6333	15056
29	imm_6_20836710	0.216	0.85	0.73	CDKAL1	6333	15056
30	rs1799964	0.209	1.19	1.42	LTA	6333	15056
31	imm_6_91029880	0.342	0.93	0.87	BACH2	6333	15056
32	imm_6_106541962	0.301	1.13	1.28	PRDM1	6333	15056
33	imm_6_159410424	0.393	1.1	1.21	TAGAP	6333	15056
34	imm_6_167326623	0.478	0.85	0.73	CCR6	6333	15056
35	imm_7_50275007	0.31	0.88	0.77	IKZF1	6333	15056
36	ccc-8-126606752-G-A	0.391	0.85	0.73	0	6333	15056
37	rs6651252	0.135	0.81	0.66	0	6333	15056
38	imm_9_4971602	0.349	1.18	1.39	JAK2	6333	15056
39	imm_9_116592706	0.318	0.83	0.68	TNFSF15	6333	15056
40	imm_9_138386317	0.411	1.18	1.39	CARD9	6333	15056
41	imm_10_6142018	0.148	0.9	0.81	IL2RA	6333	15056
42	imm_10_35575701	0.315	1.15	1.32	CREM	6333	15056
43	1kg_10_59583157	0.226	0.84	0.71	UBE2D1	6333	15056
44	imm_10_64115570	0.462	0.81	0.66	ZNF365	6333	15056
45	imm_10_80730323	0.331	0.84	0.71	ZMIZ1	6333	15056
46	imm_10_101274227	0.492	1.22	1.49	NKX2-3	6333	15056
47	rs102275	0.341	1.08	1.17	FADS1	6333	15056
48	rs694739	0.374	0.91	0.83	PRDX5	6333	15056
49	rs7927997	0.389	1.17	1.37	C11orf30	6333	15056
50	imm_12_39078567	0.025	1.74	3.03	MUC19, LRRK2	6333	15056
51	1kg_13_41950880	0.346	1.1	1.21	TNFSF11	6333	15056
52	imm_13_43355925	0.245	1.17	1.37	C13orf31	6333	15056
53	imm_14_68279952	0.416	0.93	0.87	ZFP36L1	6333	15056
54	1kg_14_87542348	0.119	1.23	1.51	GALC	6333	15056
55	imm_15_65229650	0.233	1.12	1.25	SMAD3	6333	15056
56	imm_16_28398018	0.386	1.07	1.14	IL27	6333	15056
57	NOD2 (combined rs2066844, rs2066845 and rs2066847)	0.087	2.05	14.92	NOD2	15797	22548
58	1kg_17_29617778	0.277	0.83	0.69	CCL2, CCL7	6333	15056
59	imm_17_35294289	0.458	1.14	1.3	GSMDL	6333	15056
60	imm_17_37826319	0.244	0.87	0.76	MLX	6333	15056
61	imm_18_12799340	0.153	1.25	1.56	PTPN2	6333	15056
62	1kg_19_1075835	0.247	1.16	1.35	GPX4	6333	15056
63	rs12720356	0.084	1.12	1.25	TYK2	6333	15056
64	imm_19_53906086	0.487	1.07	1.14	FUT2	6333	15056
65	rs4809330	0.291	0.89	0.8	RTEL1	6333	15056
66	imm_21_15734423	0.421	0.86	0.74	0	6333	15056
67	imm_21_44439451	0.391	1.18	1.39	ICOSLG	6333	15056
68	imm_22_20258597	0.203	1.1	1.21	YDJC	6333	15056
69	imm_22_28922487	0.471	1.08	1.17	MTMR3	6333	15056
70	imm_22_37989719	0.17	0.81	0.66	MAP3K7IP1	6333	15056

(SNP – single nucleotide polymorphism, MAF – minor allele frequency, RR_het – relative risk heterozygous for risk allele, RR_hom – relative risk homozygous for risk allele; Ncase – number of cases; Ncontrol – number of controls)

Table S2. Distribution of demographics, relationship and CD risk score by risk quartile for entire genotyped cohort

	Highest risk quartile	Third risk quartile	Second risk quartile	Lowest risk quartile
N =454	114	113	113	114
Relationship to proband				
sibling	48 (42%)	40 (35%)	46 (41%)	48 (42%)
offspring	53 (47%)	54 (48%)	50 (44%)	51 (45%)
parent	13 (11%)	19 (17%)	17 (15%)	15 (13%)
Sex				
female	72 (63%)	71 (63%)	66 (58%)	69 (61%)
male	42 (37%)	42 (37%)	47 (42%)	45 (39%)
Age				
mean	35.7	37.9	36.5	35.5
median	*36.5*	*39*	*35*	*33*
Smoking status				
current	52 (46%)	19 (17%)	5 (4%)	9 (8%)
ex	41 (36%)	27 (24%)	30 (27%)	16 (14%)
never	21 (18%)	67 (59%)	78 (69%)	89 (78%)
CD family history				
single	74 (65%)	75 (66%)	91 (81%)	90 (79%)
multiple	40 (35%)	38 (34%)	22 (19%)	24 (21%)

Table S3. Biomarker descriptive statistics for 124 individuals with complete data and variance > 0

Biomarker	**Description**	**Mean (SD) / Count**	**Median**	**Range**
VCAM	Vascular cell adhesion molecule (µg/g)	573.44 (140.77)	559.69	38.49–1034.17
ICAM	Intercellular adhesion molecule (µg/L)	239.26 (53.04)	232.26	33.54–390.27
ESEL	E-selectin (µg/L)	14.35 (4.97)	13.07	5.29–26.26
PSEL	P-selectin (µg/L)	277.70 (135.40)	231.54	134.21–729.24
LSEL	L-selectin (µg/L)	1872.42 (306.43)	1851.77	1242.69–2809.83
*ASCA IgG	Saccharomyces cerevisiae antibodies Immunoglobulin G	Negative: 112, Equivocal: 4, Positive: 8		

*ASCA IgA	Saccharomyces cerevisiae antibodies Immunoglobulin A	Negative: 113, Equivocal: 1, Positive: 10		
IL2	Interleukin 2 (ng/L)	1.17 (0.81)	0.90	0.90–6.07
*IL4	Interleukin 4 (ng/L)	2.13 (0.10)	2.12	2.12–3.18
IL6	Interleukin 6 (ng/L)	0.96 (1.28)	0.58	0.18–8.59
IL8	Interleukin 8 (ng/L)	63.87 (130.17)	4.64	0.80–415.00
*IL10	Interleukin 10 (ng/L)	0.49 (0.41)	0.37	0.37–3.75
*IL1a	Interleukin 1 alpha (ng/L)	0.26 (0.22)	0.19	0.19–2.18
IL1b	Interleukin 1 beta (ng/L)	1.18 (2.84)	0.67	0.26–31.09
IL1RA	Interleukin-1 receptor antagonist (ng/L)	321.47 (335.95)	194.64	24.30–1657.37
FC	Faecal calprotectin (µg/g)	68.58 (112.15)	36.25	10.00–866.50
VEGF	Vascular endothelial growth factor (pg/ml)	60.19 (46.44)	47.50	8.57–283.08
EGF	Endothelial growth factor (ng/L)	56.69 (39.01)	49.52	2.72–228.15
*INFy	Interferon gamma (ng/L)	0.45 (0.09)	0.44	0.44–1.20
TNFa	Tumor necrosis factor alpha (ng/L)	0.78 (0.25)	0.64	0.59–1.48
MCP1	Monocyte chemoattractant protein 1 (ng/L)	92.92 (95.69)	72.11	4.21–646.00
hs-CRP	Highly sensitive C-reactive protein (mg/L)	2.53 (3.41)	1.30	0.10–18.30

Note. Biomarker descriptive statistics for 124 individuals with complete data. * = variables with near-zero variance not included in statistical modelling

Table S4. Descriptive statistics for 124 individuals with complete data (67 high RR, 57 low RR) by Lewis score (0 = Normal, >0 = Abnormal)

Normal				
Lewis	Sex	Smoking	ASCAIgG	ASCAIgA
Normal : 98	F:67	Never :50	Negative :91	Negative :90
Abnormal: 0	M:31	Current:21	Equivocal: 2	Equivocal: 1
		Ex :27	Positive : 5	Positive : 7

Abnormal				
Lewis	Sex	Smoking	ASCAIgG	ASCAIgA
Normal : 0	F:15	Never :12	Negative :21	Negative :23
Abnormal:26	M:11	Current: 8	Equivocal: 2	Equivocal: 0
		Ex : 6	Positive : 3	Positive : 3

Normal	mean	sd	median	min	max	range	skew	kurtosis
Age	38.08	10.46	37.55	19.5	56.7	37.2	0.09	-1.16
Genetic Risk Score	0.66	1.02	0.47	-2.26	2.93	5.19	-0.24	-0.34
Number CDFDR	1.06	0.28	1	1	3	2	4.89	25.41
VCAM1	561.49	137.96	559.69	38.49	987.02	948.53	0.17	2.05
ICAM1	235.88	54.31	225.04	33.54	390.27	356.73	0.25	1.44
ESEL	13.95	4.95	12.82	5.29	26.26	20.97	0.51	-0.39
IL2	1.19	0.88	0.9	0.9	6.07	5.17	4.02	17.11
IL10	0.49	0.41	0.37	0.37	3.75	3.38	5.73	39.33
TNFa	0.78	0.26	0.6	0.59	1.48	0.89	1.17	0.05
IL1a	0.25	0.23	0.19	0.19	2.18	1.99	6.45	48.23
hs-CRP	2.24	3.32	1.1	0.1	18.3	18.2	2.91	8.89
Transit Time	228.31	76.07	232.5	66	428	362	0.38	-0.04
FC	43	48.1	25.95	10	316	306	3.25	12.69
IL6	0.93	1.34	0.55	0.18	8.59	8.41	4.01	16.93
IL8	53.22	122.81	3.99	0.8	415	414.2	2.39	4.08
IL1b	1.27	3.17	0.68	0.26	31.09	30.83	8.5	76.64
IL1RA	301.11	343.85	153.15	24.3	1657.37	1633.07	2.31	5
VEGF	60.4	48.83	46.2	8.57	283.08	274.51	2.19	5.81
EGF	56.01	38.53	49.36	2.72	228.15	225.43	1.62	4.56
MCP1	94.07	105.47	71.19	4.21	646	641.79	4.04	17.4
PSEL	272.82	136.38	226.03	134.21	729.24	595.03	1.89	2.83
LSEL	1852.8	323.24	1847.05	1242.69	2809.83	1567.14	0.52	0.1

Abnormal	mean	sd	median	min	max	range	skew	kurtosis
age	34.91	10.29	32.45	20.4	53.9	33.5	0.47	-1.05
Genetic Risk Score	1.55	1.11	1.59	-1.25	3.75	5	-0.43	0.25
Number CDFDR	1.38	0.64	1	1	3	2	1.31	0.47
VCAM1	618.51	144.8	557.43	455.43	1034.17	578.74	1.05	0.36
ICAM1	251.99	46.76	252.92	135.99	324.16	188.17	-0.42	-0.49
ESEL	15.85	4.86	15.45	7.61	25.33	17.72	0.21	-1.09
IL2	1.09	0.5	0.9	0.9	3.07	2.17	2.87	7.58

IL10	0.47	0.43	0.37	0.37	2.58	2.21	4.45	18.78
TNFa	0.78	0.21	0.74	0.59	1.21	0.62	0.87	-0.47
IL1a	0.28	0.17	0.19	0.19	0.92	0.73	2.37	5.48
hs-CRP	3.64	3.57	2.3	0.2	11.7	11.5	1	-0.21
Transit Time	236.27	87.47	225.5	89	474	385	1.1	1.58
FC	165.02	201.66	78.65	10	866.5	856.5	2.03	3.68
IL6	1.06	1.07	0.72	0.26	5.75	5.49	3.19	11.25
IL8	104.01	150.76	9.32	1.53	415	413.47	1.23	-0.08
IL1b	0.86	0.6	0.66	0.26	2.45	2.19	1.2	0.39
IL1RA	398.19	297.99	332.86	49.72	1191.84	1142.12	1.21	0.53
VEGF	59.37	36.85	52.88	14.95	167.43	152.48	1.16	1
EGF	59.26	41.43	49.55	8.69	188.13	179.44	1.43	1.92
MCP1	88.63	43.15	79.73	32.94	201.92	168.98	1.09	0.3
PSEL	296.08	132.65	259.44	136.28	636.95	500.67	0.98	-0.14
LSEL	1946.39	222.26	1947.27	1495.67	2316.14	820.47	-0.1	-1.11

Table S5. Multivariate logistic regression with stepwise selection

Variable	beta	se	z	OR	lower	upper	p
Best fit AIC = 80.35; maximal model (all predictors) AIC = 93.65							
FC	0.04	0.01	3.58	1.04	1.02	1.07	3.46E-04
Genetic Risk Score	2.42	0.75	3.21	11.22	3.25	68.41	1.35E-03
hs-CRP	0.42	0.14	2.93	1.52	1.17	2.09	3.43E-03
VEGF	-0.06	0.02	-2.66	0.94	0.89	0.98	7.80E-03
Smoking-Ex	-4.46	1.73	-2.58	0.01	0.00	0.20	9.98E-03
IL8	0.01	0.00	2.54	1.01	1.00	1.02	1.12E-02
Family burden	3.28	1.32	2.49	26.67	2.84	594.97	1.26E-02
PSEL	0.01	0.00	2.36	1.01	1.00	1.02	1.85E-02
VCAM1	0.01	0.00	2.31	1.01	1.00	1.02	2.07E-02
ESEL	0.21	0.10	2.12	1.23	1.03	1.54	3.43E-02
age	0.11	0.05	1.94	1.11	1.01	1.26	5.22E-02
ICAM1	-0.02	0.01	-1.87	0.98	0.96	1.00	6.21E-02
Transit time	0.01	0.01	1.70	1.01	1.00	1.02	8.86E-02
Smoking-Current	0.44	0.96	0.46	1.55	0.23	10.81	6.47E-01
Best fit BIC = 101.95; maximal model (all predictors) BIC = 161.34							
FC	0.01	0.00	3.57	1.01	1.01	1.02	3.63E-04
Family burden	2.22	0.73	3.04	9.24	2.29	42.65	2.37E-03
Genetic Risk Score	1.02	0.35	2.89	2.78	1.47	6.02	3.89E-03

Table S6. Confusion matrices and prediction evaluation metrics

Elastic net			**Random forest**		
Confusion Matrix and Statistics			Confusion Matrix and Statistics		
Reference			Reference		
Prediction	normal	abnormal	Prediction	normal	abnormal
normal	22	2	normal	23	1
abnormal	11	6	abnormal	10	1
Accuracy:	0.6829		Accuracy:	0.7317	
95% CI:	(0.5191, 0.8192)		95% CI:	(0.5706, 0.8578)	
No Information Rate:	0.8049		No Information Rate:	0.8049	
P-Value [Acc > NIR]:	0.9802		P-Value [Acc > NIR]:	0.91183	
Kappa:	0.2922		Kappa:	0.4011	
Mcnemar's Test P-Value:	0.0265		Mcnemar's Test P-Value:	0.01586	
Sensitivity:	0.7500		Sensitivity:	0.8750	
Specificity:	0.6667		Specificity:	0.6970	
Pos Pred Value:	0.3529		Pos Pred Value:	0.4118	
Neg Pred Value:	0.9167		Neg Pred Value:	0.9583	
Prevalence:	0.1951		Prevalence:	0.1951	
Detection Rate:	0.1463		Detection Rate:	0.1707	
Detection Prevalence:	0.4146		Detection Prevalence:	0.4146	
Balanced Accuracy:	0.7083		Balanced Accuracy:	0.7860	
Area under the curve:	0.7992		Area under the curve:	0.8712	
95% CI:	0.6191-0.9794		95% CI:	0.7465-0.9959	

Table S7. Predictor importance

	Random forest			Elastic net	
		Importance			Importance
1	FC	9.67E-02		familyBurden	0.26636
2	geneticRiskScore	2.64E-02		geneticRiskScore	0.13037
3	hs-CRP	1.51E-02		FC	0.04367
4	IL6	4.31E-03			
5	familyBurden	2.26E-03			
6	SmokingEx	4.38E-05			
7	SexM	-1.59E-04			
8	ESEL	-1.72E-04			
9	smokingCurrent	-2.68E-04			
10	EGF	-3.54E-04			
11	TNFa	-5.21E-04			
12	VCAM1	-5.94E-04			
13	IL1b	-9.66E-04			
14	IL1RA	-1.02E-03			
15	transitTime	-1.08E-03			
16	LSEL	-1.14E-03			
17	IL2	-1.29E-03			
18	VEGF	-1.32E-03			
19	ICAM1	-1.67E-03			
20	IL8	-1.72E-03			

Note. Top 20/23 predictors shown for the random forest. Random forest importance = permutation decrease in accuracy; elastic net importance = | *t*-statistic |

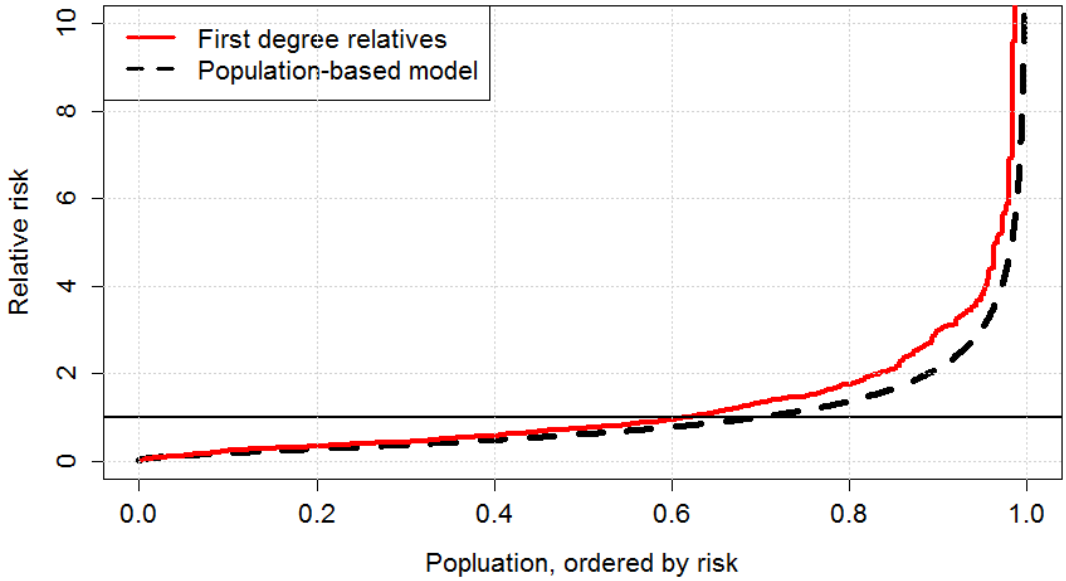
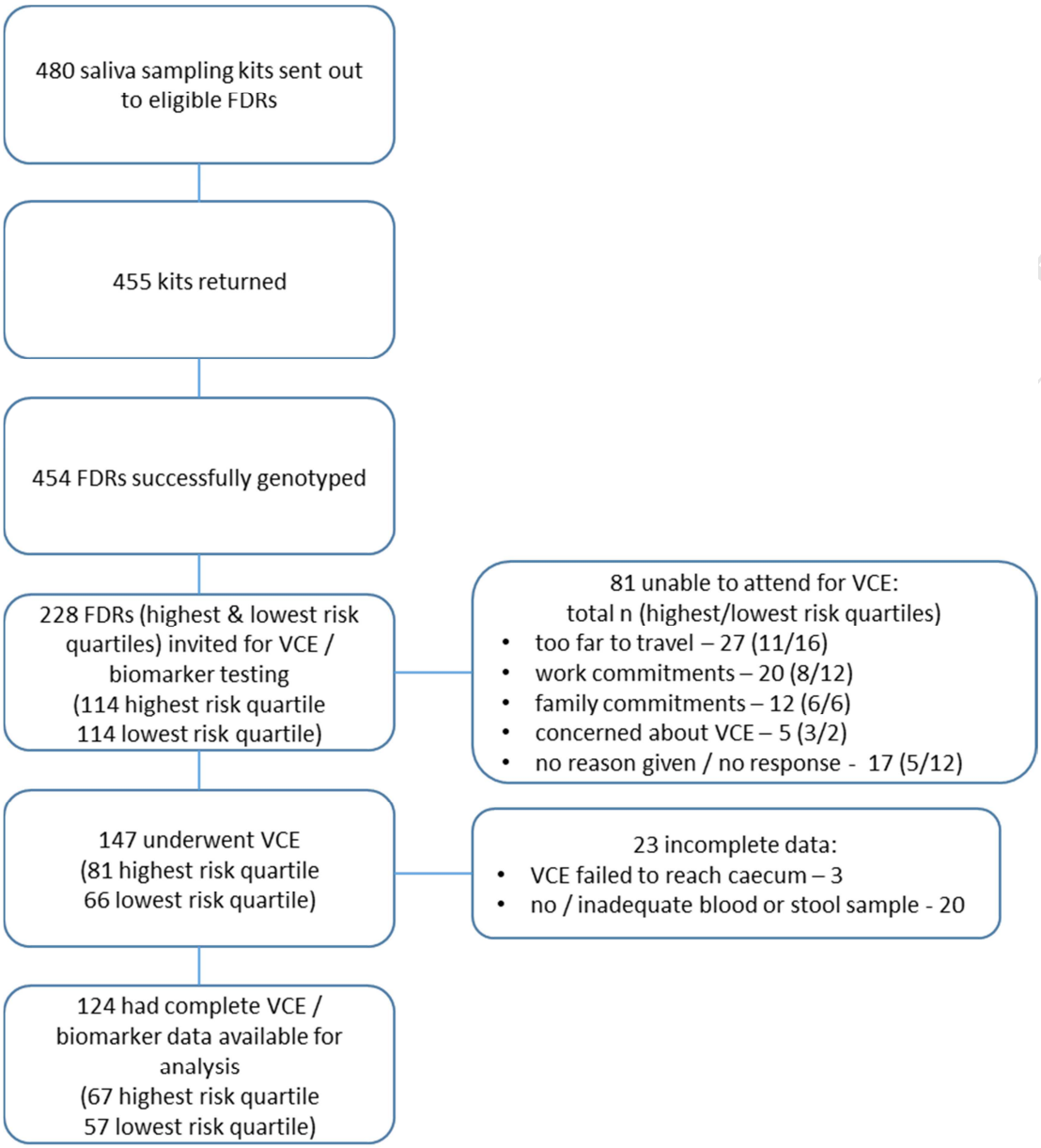


Figure S1. FDR relative risk compared to baseline population risk

Figure S2. Participant flow through the study



Note: FDR – first degree relative; VCE – video capsule endoscopy

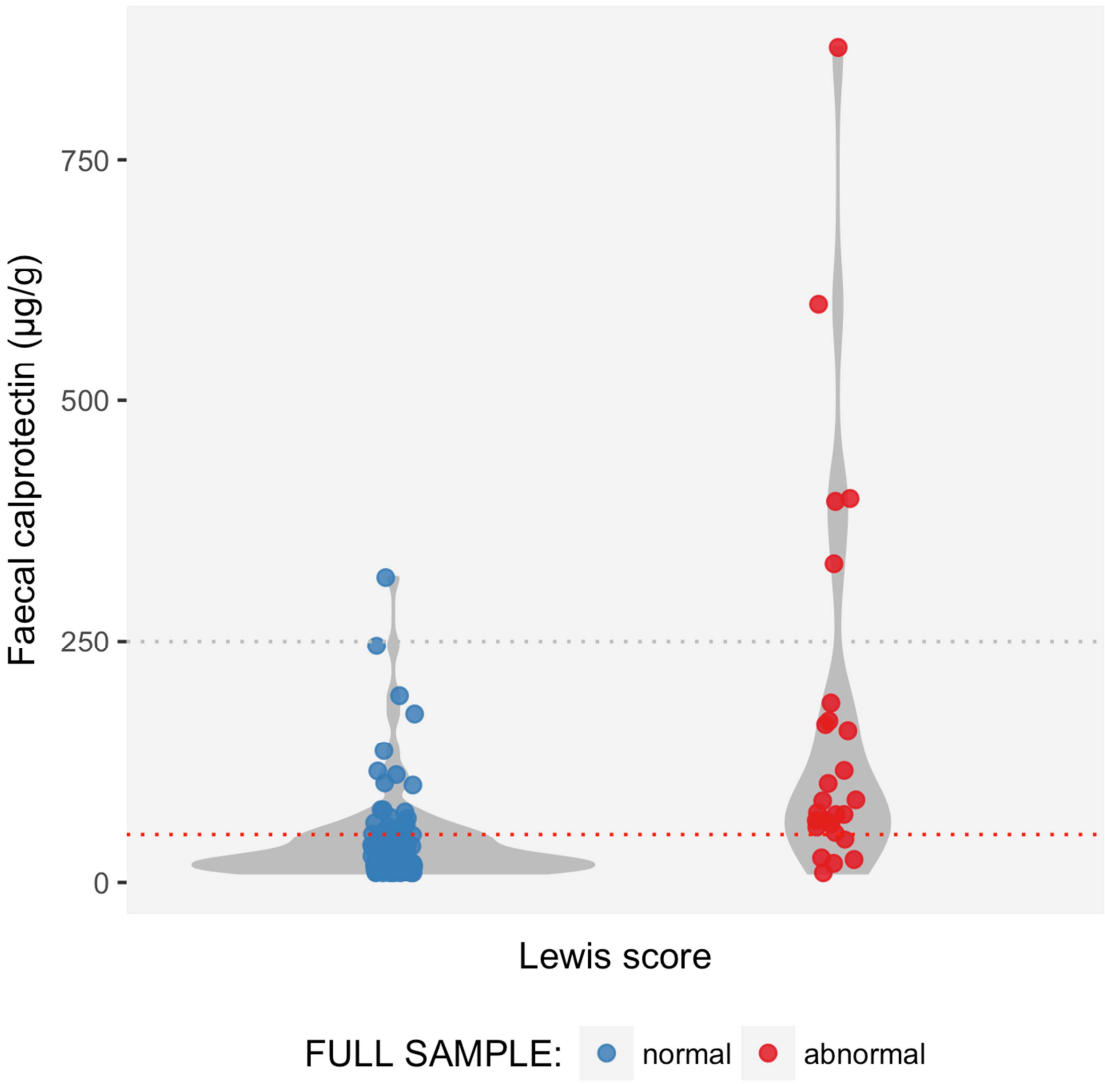


Figure S3. Faecal calprotectin levels in FDR individual with observed normal or abnormal Lewis score, where Lewis score >=790 = abnormal. Red dotted line = faecal calprotectin of 50 ug/g.

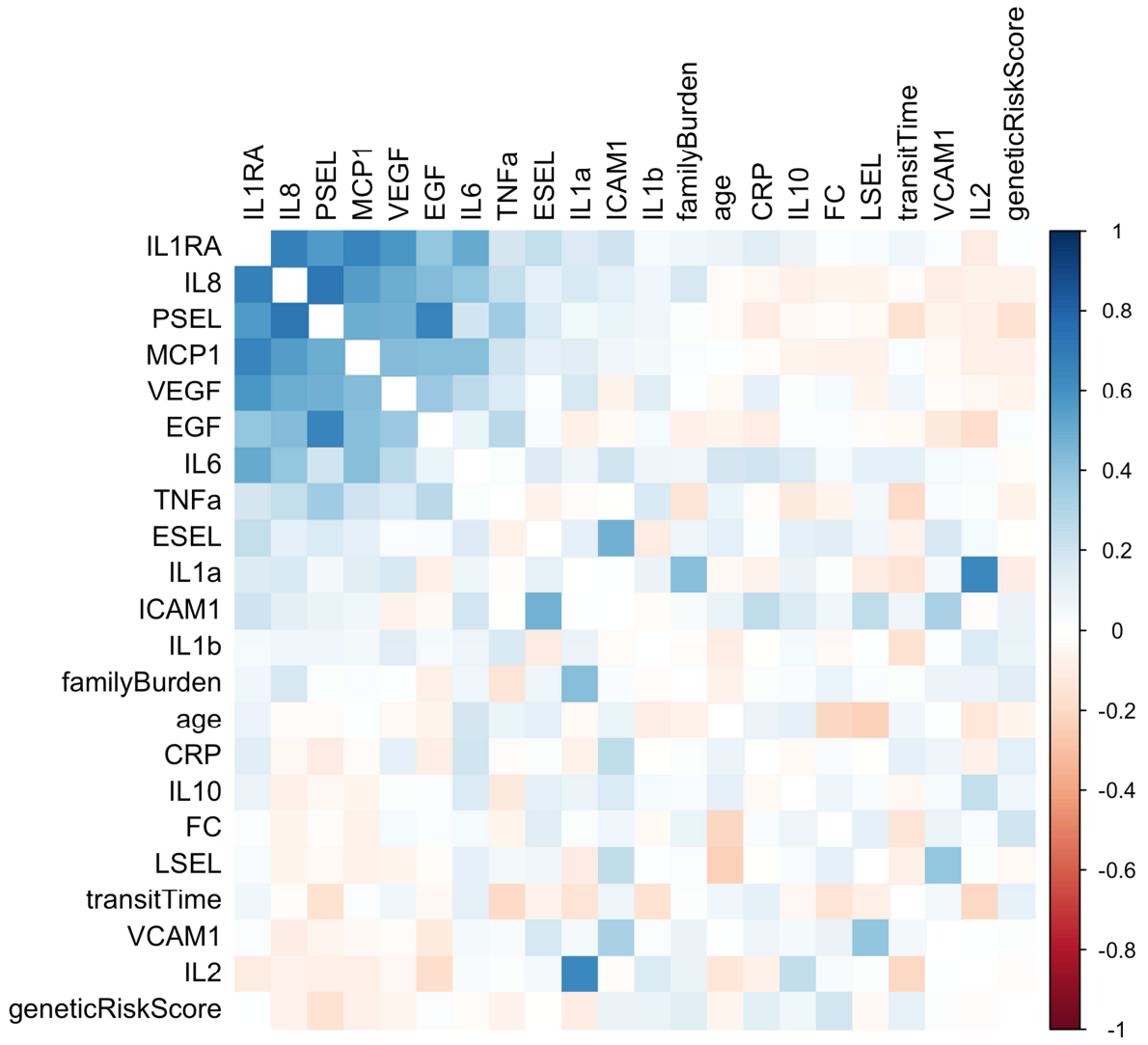


Figure S4. Correlation among continuous predictors for 124 individuals with complete data (67 high RR, 57 low RR). age = age in years; geneticRiskScore = weighted sum of risk SNPs; familyBurden = number of family members with Crohn's disease; transitTime = length of time capsule endoscopy capsule is in the gastrointestinal tract; IL2, IL10, IL1b, IL1a, IL6, IL8 = interleukin–2, –10, –1b, –1a, –6, –8; VCAM1, ICAM1 = vascular adhesion molecule, intercellular adhesion molecule; ESEL, LSEL, PSEL = E–, L–, P–selectin; EGF, VEGF = endothelial growth factor, vascular endothelial growth factor; TNFa = tumor necrosis factor alpha; MCP1 = Monocyte chemoattractant protein 1; FC = faecal calprotectin; CRP = C-reactive protein

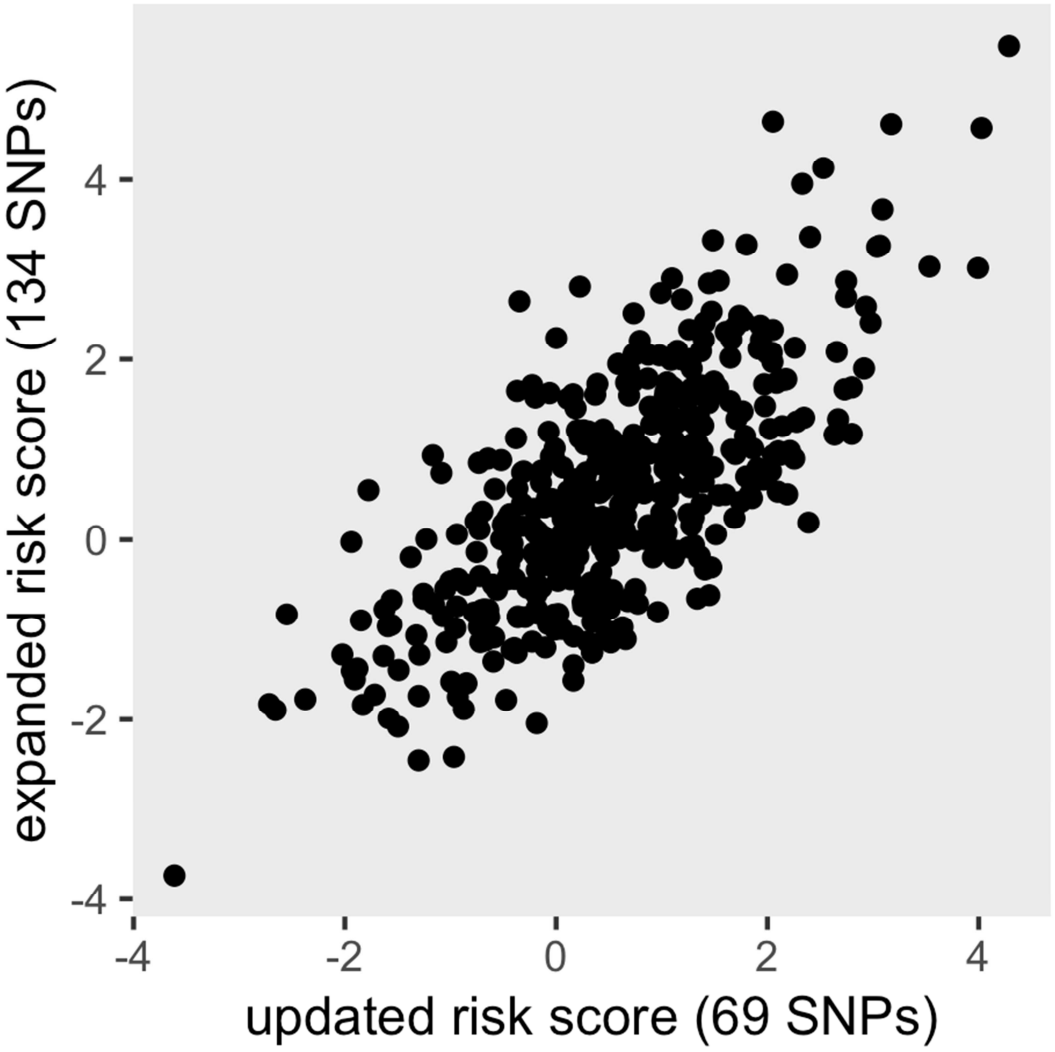


Figure S5. Plot showing correlation (r = 0.72) between the updated risk score[10] for 69 SNPs (plus *NOD2* and smoking) versus risk scores based on all 134 CD associated SNPs available in our study.